 doi:10.20944/preprints202108.0321.v1

Protein structure readouts of cancer drivers for precision medicine

Jaspreet Kaur Dhanjal[1*], Rajkumar Singh Kalra[2†*]

[1]Indraprastha Institute of Information Technology Delhi, Okhla Industrial Estate, Phase III, New Delhi 110 020, India. Electronic address: jaspreet@iiitd.ac.in

[2]AIST-INDIA DAILAB, National Institute of Advanced Industrial Science & Technology (AIST), Higashi 1-1-1, Tsukuba 305 8565, Japan. Electronic address: rajkumar-singh@oist.jp

*Correspondence: jaspreet@iiitd.ac.in (JKD); rajkumar-singh@oist.jp (RSK)

[†]Present address: Okinawa Institute of Science and Technology Graduate University, 1919-1 Tancha, Onna-son, Okinawa, 904-0495, Japan. Electronic address: rajkumar-singh@oist.jp

ABSTRACT

Cancer is fundamentally a disease of perturbed genes. Although many mutations can be marked in the genome of a cancer or transformed cell, the initiation and progression were shown to be driven by only a few mutational events *viz.* driver mutations that progressively govern and execute the functional impacts. The driver mutations are thus believed to dictate and dysregulate the subsequent cellular proliferative function/decisions thereby producing a cancerous state. Therefore, identifying the driver events from the genomic alterations in a patient's cancer cell gained large attention recently for designing better targeting therapies towards paving way for the precision cancer medicine. With rolling advancements in high-throughput omics technologies, analysis of genetic variations and gene expression profiles for cancer patients has become a routine clinical practice. However, it is anticipated that protein structural alterations resulting from such driver mutations can provide more direct and clinically relevant evidence of disease states than genetic signatures alone. This review comprehensively discusses various aspects and approaches that have been developed for the prediction of cancer drivers using genetic signatures and protein structures, and their potential application in developing precision cancer therapies.

Keywords: Cancer, mutations, cancer drivers, precision medicine, protein structure, personalized medicine, cancer therapies, genetic signatures

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 August 2021 doi:10.20944/preprints202108.0321.v1

1. INTRODUCTION

Cancer is an umbrella term used to define a group of highly heterogeneous genetic disorders. Perturbations in the DNA of the cells lead to uncontrolled proliferation and result in the development of a mass of cells. These cancer cells can further spread by gaining the capacity to migrate from the primary location of origin to distant sites in the body. Rolling advancements in next-generation sequencing technologies have increased our understanding of such genetic alterations or mutations over the past few decades. Extensive studies are been carried out to systematically characterize mutations present in the tumors of patients across different cancer types. Large-scale cancer genome projects like The Cancer Genome Atlas (TCGA) [1] and International Cancer Genome Consortium [2] have sequenced thousands of cancer genomes or whole exomes to contribute immensely to this growing knowledge base. On an average, 11000 somatic mutations have been reported per cell in a typical sporadic colorectal cancer. Similarly, different cancer cells acquire thousands of different mutations with the progression of the disease. However, the vast majority of these somatic mutations observed in different tumor or cancer cells are neutral with minimal biological effect or no significant phenotypic consequences. Such genetic alterations are termed *passenger mutations*. On the other hand, only a small fraction of these mutations, referred as *driver mutations*, are involved in the initiation of cancer by affecting the genes with critical role in maintaining or regulation the normal cell processes. In general, it includes the activation of proto-oncogenes or inactivation of tumor suppressors that imparts survival advantage to the cell despite genomic instability [3]. The minimum number of mutations can that drive a normal cell to cancerous path has remained an interesting question since long. A sincere attempt has been made in a study where the authors have applied the concept of molecular evolution to 7,664 tumors from 29 cancer types to account for the coding driver mutations. They have reported that only a small percentage of non-synonyms mutations in a cancer type constitute the driver class (1–10/tumor across tumor types). For instance, only 5% of the missense point mutations in head and neck cancers were observed to be the drivers [4].

Although we have been able to catalogue a large number of somatic mutations frequently observed in different cancer types, distinguishing between passenger and driver mutations still remains a big challenge. Finding a solution to this problem has become even more important as the cost of genome sequencing is reducing at an accelerated rate, making it possible to be routinely used in clinical practices for making therapeutic decisions as per the personalized needs of the cancer patient. Many computational algorithms have been proposed over these years to aid in the identification of cancer driver mutations. Many of these algorithms have been successful in predicting the experimentally validated driver mutations that set a mark for their performance, and have also resulted in the discovery of many new driver gene candidates adding to the existing knowledge. The rationale behind the prediction of driver mutations by these computational methods can broadly be categorized based on – (i) frequency and significance of mutations in the gene, (ii) the functional impact of the mutations in the gene, (iii) role of the gene in the network of cellular pathways, and (iv) location of the mutation in the structure of the gene product/protein.

Here we have discussed each of the above-mentioned categories of methods that is used to predict driver mutations among thousands of DNA alterations in somatic cells,

 doi:10.20944/preprints202108.0321.v1

with a special focus on gene mutations that result in structural changes in their protein product.

2. PREDICTION OF DRIVER MUTATIONS USING GENETIC SIGNATURES

2.1 Using the significance or frequency of mutations in a gene

One of the fundamental approaches used to predict driver mutations relies on the fact that each type of cancer evolves independently by the accumulation of mutations that impart some kind of survival advantage to the cells. These cells then proliferate and result in clonal expansion of the lineages and give rise to a particular oncogenic phenotype [5]. It is thus believed that the mutations that drive the cells to a particular type of cancer should occur more frequently than expected by chance across a cohort of patients. Therefore, the identification of recurring mutations can help predict driver mutations. These patterns can be observed at different levels of molecular hierarchy- single nucleotide resolution, a genetic codon, protein sequence, whole gene, or a particular cellular pathway. There are many different computational resources that employ this rationale for distinguishing driver mutations among the large pool of mutations observed in cancer samples. Some of these tools have been listed in table 1.

One of the major limitations that affect the performance of such tools is that some passenger mutations also have a high rate of occurrence but have no significant contribution in the progression of cancer, and therefore are often detected as false positives.

2.2 Assessing the functional impact of mutations in a gene

Another approach for distinguishing driver mutations from passenger mutations depends on analyzing the functional impact of mutations in the genes. These methods focus on the mutations that bring a change in amino-acid sequence of the resulting protein thereby affecting its biological activity. Therefore, here the driver mutation could be due to a single nucleotide change, the introduction of a premature stop codon, shifting of the reading frame, in-frame indels, or changes at the splice-sites. Inactivating mutations like frame-shift and nonsense mutations, in general, have a high functional impact. But single nucleotide changes often require further annotation to deduce the functional impact. Some of the popular tools that work on this rationale have also been included in table 1. Another aspect is to evaluate the positional clustering of mutations in a particular region of the gene. Mutations that lead to the activation of oncogenes or inactivation of tumor suppressors show high order of positional preferences and are known to have a high functional impact. Examples of such cancer drivers include mutations BRAF [6], KRAS [7], and TP53 [8].

In line with this, there exist many tools for the prediction of driver mutations with significant oncogenic impact, however, they need some prior information. This generally includes details about the functional domains of proteins, or regions of evolutionary conservation to decide upon the impact of mutations and distinguish between driver and passenger events.

2.3 Identifying driver modules using pathways or interaction networks

Even though identification of impactful mutations as cancer drivers is considered crucial, clinical importance can further be enhanced by finding multiples genes that work in a cohort to achieve the oncogenic phenotype. Different genes or their products

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 August 2021 doi:10.20944/preprints202108.0321.v1

interact amongst themselves to initiate signalling and carry out cellular processes or regulate various pathways. Over the past years, research in the field of cancer has characterized a number of pathways which upon perturbation by somatic mutations initiate the cancerous transformation of normal cells or help cancer cells sustain genomic instability [9]. Therefore, identification of such orchestrated key players can also yield meaningful information about cancer drivers.

Though many computational tools make use of cancer pathways or interaction networks information, there exist certain challenges that need to be addressed for improving the performance of such tools. Complex gene interacting networks possess difficulty in assessing statistically significant patterns of recurrence for identifying the cancer-driving events. Not all the genes involved in the deregulation of a particular cancer pathway are mutated at a constant rate. Detection of genes with low mutational frequency is always difficult with a clear statistical significance.

3. PREDICTION OF DRIVER MUTATIONS USING PROTEIN STRUCTURAL INFORMATION

Instead of evaluating the functional impact of mutations by assessing the mutational clusters in the gene/protein sequence, the other effective approach is to focus on the mutational hotspots based on spatial proximity in the 3D structure of the resultant protein [10]. Certain types of mutations, for instance, gain-of-function mutations in proto-oncogenes are crowded in particular sections of the protein and significantly contribute to its function in the initiation and progression of cancer. Other mutations may disrupt the structural stability of the proteins or interrupt their interaction with other proteins, DNA, small ligands, or other biological molecules. The methods that detect cancer driver genes with a large bias in clustering mutations may skip cancer drivers whose mutations are distributed far apart across the entire gene or protein sequence but lie close to each other in 3D space. Thus, enrichment analysis of mutations by mapping them to protein structures becomes important [11]. These structure-based methods compute distances between amino acid residues and generate residue-to-residue contact networks by deriving information from the 3D structures of the proteins, and locate groups of spatially proximal mutated residues as hotspots. Further, the use of structural information can also help in the detection of mutational clusters in post-translationally modified sites that usually include phosphorylation, acetylation and ubiquitination sites [12]. The functional effect of missense mutations in the protein-ligand binding sites/orthosteric sites can also be easily detected using this approach. This can further be extended to mutations in regulatory allosteric sites, derived from the 3D protein structures, located away from the ligand-binding site to detect potentially impactful driver mutations in cancer patients [12]. Examples of computational resources that can be included in this category are listed in table 1.

Although these methods that apprehend the structural consequences of gene mutations perform well, they fail to capture the protein dynamics. Proteins are dynamic in nature and can take multiple conformations [13]. Many biophysical studies have reported the effect of protein motion on change in its functionality [14]. Using the crystal structure of the protein reflects information pertaining to only one static snapshot likely from the bottom of the free-energy landscape. However, a protein can dynamically oscillate in low-energy states at the bottom of the free-energy funnel or even undergo major conformational changes as a result of genetic alterations. Therefore, efforts are being made to improve the sensitivity of these structure-based computations by incorporating the protein dynamics in the framework [14c].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 August 2021 doi:10.20944/preprints202108.0321.v1

Despite the immense potential, the incomplete structural coverage of the proteome is a hurdle for these structure-based computational resources.

Table 1. Computational resources for the prediction of cancer driver mutations

S. No.	Resource	Accessibility	Reference
Using the significance or frequency of mutations in a gene			
1.	MutSigCV	https://www.genepattern.org/modules/docs/mutsigcv	[15]
Assessing the functional impact of mutations in a gene			
2.	OncodriveFM	http://bg.upf.edu/group/projects/oncodrive-fm.php	[16]
3.	OncodriveFML	http://bbglab.irbbarcelona.org/oncodrivefml/home	[17]
4.	DriverML	https://github.com/HelloYiHan/DriverML	[18]
5.	OncodriveCLUST	http://bg.upf.edu/group/projects/oncodrive-clust.php	[19]
Using pathways or interaction networks			
6.	IntOGen-mutations	https://www.intogen.org	[20]
7.	PathScan	http://genome.wustl.edu/software/pathscan	[21]
8.	CBNA	https://github.com/pvvhoang/CancerDriver	[22]
9.	DriverNet	http://compbio.bccrc.ca/software/drivernet	[23]
10.	Sakoparnig et al.	http://www.cbg.ethz.ch/software/mutationtiming	[24]
11.	TieDIE	https://sysbiowiki.soe.ucsc.edu/tiedie	[25]
12.	CICERO	https://github.com/stjude/Cicero	[26]
Using protein structural information			
13.	ActiveDriver	https://cran.r-project.org/web/packages/ActiveDriver/ActiveDriver.pdf	[27]
14.	SGDriver	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4739678/	[28]
15.	Allodriver	http://mdl.shsmu.edu.cn/ALD/	[29]
16.	HotMAPS	https://github.com/karchinlab/HotMAPS	[30]
17.	3DHotSpots	https://github.com/taylor-lab/hotspots/blob/master/LINK_TO_MUTATIONAL_DATA	[31]
18.	HotSpot3D	https://github.com/ding-lab/hotspot3d	[32]
19.	e-Driver	https://github.com/eduardporta/ e-Driver.git	[33]
20.	DUET	http://structure.bioc.cam.ac.uk/duet	
Using dynamic structural information of proteins			
21.	Kumar et al.	https://github.com/gersteinlab/HotCommics	[14c]
22.	Sayilgan et al.	Paper link: https://pubmed.ncbi.nlm.nih.gov/33550612/	[34]

4. PREDICTION OF DRIVER MUTATIONS FOR INDIVIDUAL PATIENTS

The methods discussed so far are based on analysis of genomic data at the macro or population level. However, different patients suffering from the same cancer type are also known to possess different alterations at genomic level despite similar oncogenic phenotypes. This can be attributed to the presence of different driver mutations that led to similar end results. Hence it becomes important to investigate mutations or genomic signatures in each and every individual to find their genome-specific driver mutations and prescribe therapies according to their personal needs. Most of the methods in this category try to connect the mutated genes/or gene products with their interacting partners using transcriptomic data and pathway knowledge base. The idea is to look for the mutations that have higher connectivity in these interaction networks, and therefore can have a significant biological impact. The most impactful mutation is the one that has higher network connectivity and affects the expression of a greater number of downstream genes. The three tools that can be covered under this category have been listed in table 2.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 August 2021 doi:10.20944/preprints202108.0321.v1

Table 2. Prediction of driver mutations for individuals based on their genomic signatures.

S. No.	Resource	Accessibility	Reference
1.	DawnRank	http://bioen-compbio.bioen.illinois.edu/DawnRank/	[35]
2.	SCS	http://sysbio.sibcb.ac.cn/cb/chenlab/software.htm	[36]
3.	PNC	https://github.com/NWPU-903PR/PNC	[37]

5. CONCLUSION

It's difficult to mark an oncogene or tumor suppressor gene that is uniformly deleted or activated across cancers. Cancers originated from a common tissue rarely exhibit uniform genetic mutations, yet these tumor types can share mutations in specific genes or distinct genes that share a common growth-regulatory pathway. Advances in genome sequencing made it possible to scan and annotate specific deletions in different cancers and help oncologists to catalogue driver and passenger mutation events. This practice, distinguishing mutation-types strengthened its applications in clinical practices towards making therapeutic decisions as per the personalized needs of the cancer patient. These approaches largely exploited the utility of computational algorithms in predicting the valid driver mutations that set a mark for their performance and pave the path to the discovery of new driver events in diverse cancers. In this report, we comprehensively discussed the rationale to predict the driver mutations broadly, based on the- (i) frequency and significance of mutations, and, (ii) the functional impact of the mutations in the gene, (iii) role of the gene in the network of cellular pathways, and (iv) location of the mutation in the structure of the gene product/protein. We summarized available tools and knowledge generated on these aspects and reported the advantages and limitations of these methods/approaches. With an emphasis on the prediction of driver mutations using protein structural information, we further reviewed their impact due to spatial proximity in crucial regions of protein, post-translationally modified sites, and protein-ligand binding sites/orthosteric sites. Moreover, we underlined the importance of incorporating the protein dynamics in the framework to further improve the sensitivity of structure-based computations in finding the impact of driver mutations. Conclusively, the present report comprehensively shed light on the various aspects and approaches presently being employed in the prediction of the cancer drivers and their potential application in developing precision cancer therapies.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 August 2021 doi:10.20944/preprints202108.0321.v1

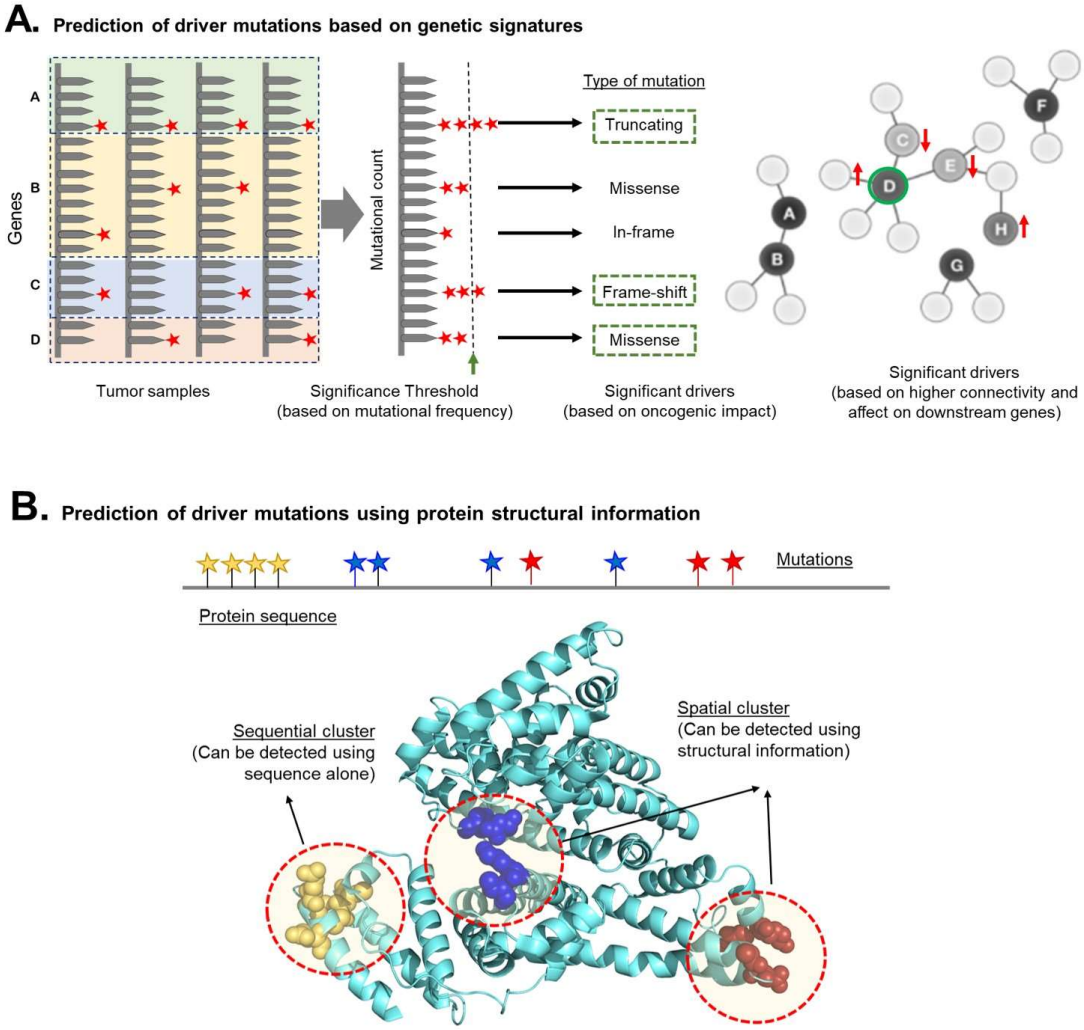

Figure 1. Schematic diagram showing prediction of driver mutations based on genetic signatures (A), and using protein structural information (B).

CONCENT FOR PUBLICATION.

Not applicable.

REFERENCES

This study is not supported by any funding of a scientific agency to disclose.

CONFLICT OF INTEREST

The authors declare no conflict of interest, financial or otherwise.

ACKNOWLEDGEMENTS

N/A.

REFERENCES

[1] Chang, K.; Creighton, C.; Davis, C.; Donehower, L. J. N. G. *Nature Genetics* **2013,** *45* (10), 1113-1120.

[2] Consortium, I. C. G. *Nature* **2010,** *464* (7291), 993.

[3] Martincorena, I.; Campbell, P. J. *Science* **2015,** *349* (6255), 1483-1489.

[4] Martincorena, I.; Raine, K. M.; Gerstung, M.; Dawson, K. J.; Haase, K.; Van Loo, P.; Davies, H.; Stratton, M. R.; Campbell, P. J. *Cell* **2017,** *171* (5), 1029-1041. e21.

[5] Nowell, P. C. J. S. *Science* **1976,** *194* (4260), 23-28.

[6] Davies, H.; Bignell, G. R.; Cox, C.; Stephens, P.; Edkins, S.; Clegg, S.; Teague, J.; Woffendin, H.; Garnett, M. J.; Bottomley, W. *Nature* **2002,** *417* (6892), 949-954.

[7] Bos, J. L. *Mutation Research* **1988,** *195* (3), 255-271.

[8] Olivier, M.; Hollstein, M.; Hainaut, P. *Cold Spring Harbor Perspectives in Biology* **2010,** *2* (1), a001008.

[9] Hanahan, D.; Weinberg, R. A. *Cell* **2011,** *144* (5), 646-674.

[10] Miller, M. L.; Reznik, E.; Gauthier, N. P.; Aksoy, B. A.; Korkut, A.; Gao, J.; Ciriello, G.; Schultz, N.; Sander, C. *Cell Systems* **2015,** *1* (3), 197-209.

[11] Ryslik, G. A.; Cheng, Y.; Modis, Y.; Zhao, H. *BMC Bioinformatics* **2016,** *17* (1), 1-13.

[12] Pham, V. V. H.; Liu, L.; Bracken, C.; Goodall, G.; Li, J.; Le, T. D. *Theranostics* **2021,** *11* (11), 5553.

[13] Tsai, C.-J.; Nussinov, R. *Physical Chemistry Chemical Physics* **2014,** *16* (14), 6332-6341.

[14] (a) Henzler-Wildman, K.; Kern, D. *Nature* **2007,** *450* (7172), 964-972; (b) Mitternacht, S.; Berezovsky, I. N. *PLoS Computational Biology* **2011,** *7* (9), e1002148; (c) Kumar, S.; Clarke, D.; Gerstein, M. B. *Proceedings of the National Academy of Sciences* **2019,** *116* (38), 18962-18970.

[15] Lawrence, M. S.; Stojanov, P.; Polak, P.; Kryukov, G. V.; Cibulskis, K.; Sivachenko, A.; Carter, S. L.; Stewart, C.; Mermel, C. H.; Roberts, S. A. *Nature* **2013,** *499* (7457), 214-218.

[16] Gonzalez-Perez, A.; Lopez-Bigas, N. *Nucleic Acids Research* **2012,** *40* (21), e169-e169.

[17] Mularoni, L.; Sabarinathan, R.; Deu-Pons, J.; Gonzalez-Perez, A.; López-Bigas, N. *Genome Biology* **2016,** *17* (1), 1-13.

[18] Han, Y.; Yang, J.; Qian, X.; Cheng, W.-C.; Liu, S.-H.; Hua, X.; Zhou, L.; Yang, Y.; Wu, Q.; Liu, P. *Nucleic Acids Research* **2019,** *47* (8), e45-e45.

[19] Tamborero, D.; Gonzalez-Perez, A.; Lopez-Bigas, N. *Bioinformatics* **2013,** *29* (18), 2238-2244.

[20] Gonzalez-Perez, A.; Perez-Llamas, C.; Deu-Pons, J.; Tamborero, D.; Schroeder, M. P.; Jene-Sanz, A.; Santos, A.; Lopez-Bigas, N. *Nature Methods* **2013,** *10* (11), 1081-1082.

[21] Wendl, M. C.; Wallis, J. W.; Lin, L.; Kandoth, C.; Mardis, E. R.; Wilson, R. K.; Ding, L. *Bioinformatics* **2011,** *27* (12), 1595-1602.

[22] Pham, V. V.; Liu, L.; Bracken, C. P.; Goodall, G. J.; Long, Q.; Li, J.; Le, T. D. *PLoS Computational Biology* **2019,** *15* (12), e1007538.

[23] Bashashati, A.; Haffari, G.; Ding, J.; Ha, G.; Lui, K.; Rosner, J.; Huntsman, D. G.; Caldas, C.; Aparicio, S. A.; Shah, S. P. *Genome Biology* **2012,** *13* (12), 1-14.

[24] Sakoparnig, T.; Fried, P.; Beerenwinkel, N. *PLoS Computational Biology* **2015,** *11* (1), e1004027.

[25] Paull, E. O.; Carlin, D. E.; Niepel, M.; Sorger, P. K.; Haussler, D.; Stuart, J. M. *Bioinformatics* **2013,** *29* (21), 2757-2764.

[26] Tian, L.; Li, Y.; Edmonson, M. N.; Zhou, X.; Newman, S.; McLeod, C.; Thrasher, A.; Liu, Y.; Tang, B.; Rusch, M. C. *Genome Biology* **2020,** *21*, 1-18.

[27] Reimand, J.; Bader, G. D. *Molecular Systems Biology* **2014,** *10* (8).

[28] Zhao, J.; Cheng, F.; Wang, Y.; Arteaga, C. L.; Zhao, Z. *Molecular Cellular Proteomics* **2016,** *15* (2), 642-656.

[29] Song, K.; Li, Q.; Gao, W.; Lu, S.; Shen, Q.; Liu, X.; Wu, Y.; Wang, B.; Lin, H.; Chen, G. *Nucleic Acids Research* **2019,** *47* (W1), W315-W321.

[30] Tokheim, C.; Bhattacharya, R.; Niknafs, N.; Gygax, D. M.; Kim, R.; Ryan, M.; Masica, D. L.; Karchin, R. *Cancer Research* **2016,** *76* (13), 3719-3731.

[31] Gao, J.; Chang, M. T.; Johnsen, H. C.; Gao, S. P.; Sylvester, B. E.; Sumer, S. O.; Zhang, H.; Solit, D. B.; Taylor, B. S.; Schultz, N. *Genome Medicine* **2017,** *9* (1), 1-13.

[32] Niu, B.; Scott, A. D.; Sengupta, S.; Bailey, M. H.; Batra, P.; Ning, J.; Wyczalkowski, M. A.; Liang, W.-W.; Zhang, Q.; McLellan, M. D. *Nature Genetics* **2016,** *48* (8), 827-837.

[33] Porta-Pardo, E.; Godzik, A. *Bioinformatics* **2014,** *30* (21), 3109-3114.

[34] Sayılgan, J. F.; Haliloğlu, T.; Gönen, M. *Proteins: Structure, Function, Bioinformatics* **2021,** *89* (6), 721-730.

[35] Hou, J. P.; Ma, J. *Genome Medicine* **2014,** *6* (7), 1-16.

[36] Guo, W.-F.; Zhang, S.-W.; Liu, L.-L.; Liu, F.; Shi, Q.-Q.; Zhang, L.; Tang, Y.; Zeng, T.; Chen, L. *Bioinformatics* **2018,** *34* (11), 1893-1903.

[37] Guo, W.-F.; Zhang, S.-W.; Zeng, T.; Li, Y.; Gao, J.; Chen, L. *PLoS Computational Biology* **2019,** *15* (11), e1007520.